CLAIMS:

1. A method of text clustering for the generation of language models, a text (300) featuring a plurality of text units (320, 322,...), each of which having at least one word (302, 304,...), the method of text clustering comprising the steps of:

- assigning each of the text units (320, 322,...) to one of a plurality of provided clusters (330, 332,...),
- determining for each text unit a set of emission probabilities (340, 350), each emission probability (342, 344,...,352, 354,...) being indicative of a correlation between the text unit (320, 322,...) and a cluster (330, 332,...), the set of emission probabilities being indicative of the correlations between the text unit and the plurality of clusters,
- determining a transition probability (362, 364,...) being indicative that a first cluster (330) being assigned to a first text unit (320) in the text is followed by a second cluster (332) being assigned to a second text unit (322) in the text, the second text unit (322) subsequently following the first text unit (320) within the text,
- performing an optimization procedure based on the emission probability and the transition probability in order to assign each text unit to a cluster.

2. The method according to claim 1, wherein the optimization procedure comprises evaluating a target function by making use of statistical parameters based on the emission and transition probability, the statistical parameters comprising word counts, transition counts, cluster sizes and cluster frequencies.

3. The method according to claim 2, wherein the optimization procedure comprises a re-clustering procedure, the re-clustering procedure comprising the steps of:

(a) performing a modification by assigning a first text unit (320) that has been assigned to a first cluster (330) to a second cluster (332),

(b) evaluating the target function by making use of the statistical parameters accounting for the performed modification,

(c) assigning the text unit (320) to the second cluster (332) when the result of the target function has improved compared to the corresponding result based on the first text unit (320) being assigned to the first cluster (330),

(d) repeating steps (a) through (c) for each of the plurality of clusters (330, 332, ...) being the second cluster,

(e) repeating steps (a) through (d), for each of the plurality of text units (320, 322,...) being the first text unit.

4. The method according to claim 2 or 3, wherein a smoothing procedure is applied to the target function, the smoothing procedure comprising a discount technique, a backing-off technique, or an add-one smoothing technique.

5. The method according to any one of the claims 1 to 4, comprising a weighting functionality in order to decrease or increase the impact of the transition or emission probability on the target function.

6. The method according to claim 4 or 5, wherein the smoothing procedure further comprises an add-x smoothing technique making use of adding a number x to the word counts and adding a number y to the transition counts in order to modify the smoothing procedure and/or the weighting functionality.

7. The method according to any one of the claims 2 to 6, wherein evaluating of the target function further comprises making use of modified emission (340, 350) and transitions probabilities (360) in form of a leaving-one-out technique.

8. The method according to any one of the claims 1 to 7, wherein a text unit (320) either comprises a single word (302), a set of words (302, 304,...), a sentence or a set of sentences.

9. The method according to any one of the claims 1 to 8, wherein the number of clusters (330, 332,...) does not exceed a predefined maximum number of clusters.

10. The method according to any one of the claims 1 to 9, wherein the text (300) comprises a weakly annotated structure with a number of labels assigned to at least one text unit (320) or to a set of text units (320, 322,...), the method of text clustering further comprising assigning the same cluster to text units having assigned the same label.

11. A computer program product for text clustering for the generation of language models, a text (300) featuring a plurality of text units (320, 322,...), each of which having at least one word (302, 304,...), the computer program product comprising program means for:

- assigning each of the text units (320, 322,...) to one of a plurality of provided clusters (330, 332,...),
- determining for each text unit a set of emission probabilities (340, 350), each emission probability (342, 344,..., 352, 354,...) being indicative of a correlation between the text unit (320, 322,...) and a cluster (330, 332,...), the set of emission probabilities being indicative of the correlations between the text unit and the plurality of clusters,

- determining a transition probability (362, 364,...) being indicative that a first cluster (330) being assigned to a first text unit (320) in the text is followed by a second cluster (332) being assigned to a second text unit (322) in the text, the second text unit (322) subsequently following the first text unit (320) within the text,
- performing an optimization procedure based on the emission probability and the transition probability in order to assign each text unit to a cluster.

12. The computer program product according to claim 11, wherein the program means for performing the optimization procedure further comprise evaluating a target function by making use of statistical parameters based on the emission and transition probability, the statistical parameters comprising word counts, transition counts, cluster sizes and cluster frequencies.

13. The computer program product according to claim 11, wherein the program means for performing the optimization procedure further comprise program means for re-clustering, the re-clustering program means are adapted to perform the steps of:

(a) performing a modification by assigning a first text unit (320) that has been assigned to a first cluster (330) to a second cluster (332),

(b) evaluating the target function by making use of the statistical parameters accounting for the performed modification,

(c) assigning the text unit (320) to the second cluster (332) when the result of the target function has improved compared to the corresponding result based on the first text (320) unit being assigned to the first cluster (330),

(d) repeating steps (a) through (c) for each of the plurality of clusters (330, 332,...) being the second cluster,

(e) repeating steps (a) through (d), for each of the plurality of text units (320, 322,...) being the first text unit.

14. The computer program product according to claim 12 or 13, further comprising program means being adapted to perform a smoothing procedure for the target function, the smoothing procedure comprising a discount technique, a backing-off technique, an add-one smoothing technique or separate add-x and add-y smoothing techniques for the word and cluster transition counts.

15. The computer program product according to any one of the claims 11 to 14, further comprising program means providing a weighting functionality in order to decrease or increase the impact of the transition or emission probability on the target function.

16. The computer program product according to any one of the claims 11 to 15, wherein a text unit (320) either comprises a single word (302), a set of words (302, 304,...), a sentence or a set of sentences.

17. A text clustering system for the generation of language models, a text (300) featuring a plurality of text units (320, 322,...), each of which having at least one word (302, 304,...), the text clustering system comprising:

- means for assigning each of the text units (320, 322,...) to one of a plurality of provided clusters (330, 332,...),
- means for determining for each text unit a set of emission probabilities (340, 350), each emission probability (342, 344,..., 352, 354) being indicative of a correlation between the text unit (320, 322,...) and a cluster (330, 332,...), the set of emission probabilities being indicative of the correlations between the text unit and the plurality of clusters,
- means for determining a transition probability (362, 364,...) being indicative that a first cluster (330) being assigned to a first text unit (320) in the text is followed by a second cluster (332) being assigned to a second text unit (322) in the text, the second text unit (322) subsequently following the first text unit (320) within the text,

- means for performing an optimization procedure based on the emission probability and the transition probability in order to assign each text unit to a cluster.

18. The text clustering system according to claim 17, wherein means for performing the optimization procedure are adapted to evaluate a target function and to perform a re-clustering procedure by making use of statistical parameters based on the emission and transition probability, the statistical parameters comprising word counts, transition counts, cluster sizes and cluster frequencies comprises a re-clustering procedure, the re-clustering procedure comprising the steps of:

(a) performing a modification by assigning a first text unit (320) that has been assigned to a first cluster (330) to a second cluster (332),

(b) evaluating the target function by making use of the statistical parameters accounting for the performed modification,

(c) assigning the text unit (320) to the second cluster (332) when the result of the target function has improved compared to the corresponding result based on the first text unit (320) being assigned to the first cluster (330),

(d) repeating steps (a) through (c) for each of the plurality of clusters (330, 332,...) being the second cluster,

(e) repeating steps (a) through (d), for each of the plurality of text units (320, 322,...) being the first text unit.

19. The text clustering system according to claim 18, further comprising means being adapted to apply a smoothing procedure to the target function, the smoothing procedure comprising a discount technique, a backing-off technique, an add-one smoothing technique or separate add-x and add-y smoothing techniques for the word and cluster transition counts.

20. The text clustering system according to any one of the claims 17 to 19, wherein a text unit (320) can either comprise a single word (302), a set of words (302,

304,...), a sentence or a set of sentences, the clustering further comprising means being adapted to provide a weighting functionality in order to decrease or increase the impact of the transition and emission probability on the target function.